CS 134: Networks	Lecture 7, Monday, 2/13
Prof. Yaron Singer	Spring 2017

1 Overview

Question: *how do natural sample biases shape our perception of the world?*

1.1 Example: Everyone at the Gym Looks Fitter than Me

The gym paradox: *The average fitness of the people you see in the gym is higher than the average fitness of the gym's members.*

To see this, consider a gym with five members. Suppose member $n \in \{0, 1, 2, 3, 4\}$ spends n hours a week in the gym, and his fitness level is n. Suppose these hours are distributed uniformly at random, so that the probability you sample someone at the gym is proportional to the number of hours someone spends there. Let p_n be the probability of sampling member n. By assumption $p_n = kn$, where k is some constant. Since probabilities must add up to 1, we compute $k = \frac{1}{10}$. What is the expected fitness of someone you see at the gym? It is

$$0 \cdot \frac{0}{10} + 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{3}{10} = 3$$

So even though the average fitness of this gym members is 2, the expected fitness of a person randomly sampled at the gym is 3 (i.e., the time-weighted average). No wonder most people think they are less fit than the average gym member.

- What is the average fitness of gym members?
- What is the average fitness of someone at the gym when I visit?

1.2 Example: I am always in classes bigger than average

The class size paradox: *Students experience an average class size larger than the school's average class size.*

To clarify this statement and provide some intuition, consider the following example. Consider a department with two classes; every student in the department takes one or the other, but not both. One is a large introductory course with 90 freshmen in it. The other is an advanced seminar with 10 seniors.

- The average class size of this department's lectures is 50
- The average of the class-size experiences of this department's students is 82 – 10 student's experiences are a class size of 10 and 90 student's experiences are a class size of 90; the average: $\frac{10*10+90*90}{100}$.

Note that the average class size is an unweighted average, whereas the average of the class-size experiences is a weighted average where experiences of more crowded classes have a higher weight.

Intuition: the variation in class size has the consequence that only a few students at a time can experience the smaller classes while very many students can simultaneously experience the larger classes. This drives up the average experienced class size above the average class size.

In this lecture, we will focus on a particular class size phenomenon: the friendship paradox.

2 The Friendship Paradox: Why Your Friends Have More Friends than You

Now that you have a sense of what selection bias can do, we'll talk about an application in graphs. The friendship paradox is the fact that most people's friends are typically more popular than they are (for example, this was true for more than 90% of Facebook users (in 2011)). This is not really a fact about people, but a special case of the "class size paradox". Studying it carefully will illustrate how to use the words we've learned.

Unlike our previous two examples, the friendship paradox is explicitly about graphs. And it will be our first theorem. A story (theorem) in graph theory has "topic sentence" (statement of the theorem) that says what the *point* or *takeaway* of the story will be. You can think of this as the main claim in a persuasive essay. It's a summary of what we're saying. Sometimes we state it first and then argue for ("prove") it. Other times, we figure something out and summarize our conclusion as a theorem statement at the end.

Two Distributions: Friends of Individuals and Friends of Friends. The phenomenon of people finding that their friends have more friends than they do can be partially understood by recognizing the difference between the distribution of numbers of friends of individuals and the distribution of the numbers of friends of friends. The distribution of friends of individuals is just the usual distribution of numbers of friends that we would usually examine, but the distribution of friends of friends includes some of the same individuals over and over. When each individual compares him- or herself with the average number of friends of his or her friends, the comparison is with a sample from the numbers of friends of friends, which is a different distribution from that of numbers of friends among individuals.

The formal statement of the friendship paradox is the following:

Theorem (The Friendship Paradox)**.** *In any graph, the average degree of nodes is no more than the average degree of all the neighbors.*

Proof. First, let's calculate the average degree of nodes. The average degree is simply the sum of all degrees divided by the number of nodes. For every node i in the graph, let d_i denote its degree, and for a graph with n nodes we have that the average degree μ is:

$$\mu = \frac{\sum_{i=1}^{n} d_i}{n}$$

Now, let's calculate the average degree of neighbors. Here, we want to count the total degree of and divide it by the number of neighbors. For every node:

$$\frac{\sum_{i=1}^{n} \text{total degree of neighbors of i}}{\text{total number of neighbors in the graph}}$$

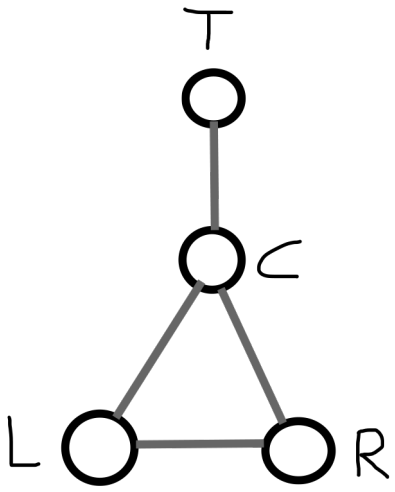
For every node i, its degree d_i is added to every one of its d_i neighbors. Hence the total degree of neighbors is $\sum_i d_i^2$ and averaging this over the total number of neighbors we get that the average degree over all neighbors is $\sum_i d_i^2 / \sum_i d_i$. To see how this compares to the average degree of nodes we can manipulation this term:

$$\begin{aligned}
\frac{\sum_i d_i^2}{\sum_i d_i} =& \frac{\sum_i d_i^2}{\sum_i d_i} - \frac{\sum_i d_i}{n} + \frac{\sum_i d_i}{n} \\
=& \frac{\frac{\sum_i d_i^2}{n}}{\frac{\sum_i d_i}{n}} - \frac{\left(\frac{\sum_i d_i}{n}\right)^2}{\frac{\sum_i d_i}{n}} + \frac{\sum_i d_i}{n} \\
=& \frac{\frac{\sum_i d_i^2}{n} - \left(\frac{\sum_i d_i}{n}\right)^2}{\frac{\sum_i d_i}{n}} + \frac{\sum_i d_i}{n} \\
=& \frac{\sigma^2}{\mu} + \mu
\end{aligned}$$

Where σ^2 denotes the variance[1] □

The term "the friendship paradox" and the above proof are due to Scott Feld [1]. Ideally, to make a statement like "your friends have more friends than you", we would like to prove something a bit stronger, namely that for every node its degree is bounded from above by the average degree of its neighbors.

Small example. Consider the following network:



- The average degree of individuals is $\frac{1+3+2+2}{4} = 2$.
- The average degree of friends is $\frac{1+3*3+2*2+2*2}{8} = \frac{9}{4}$. To see this, note that

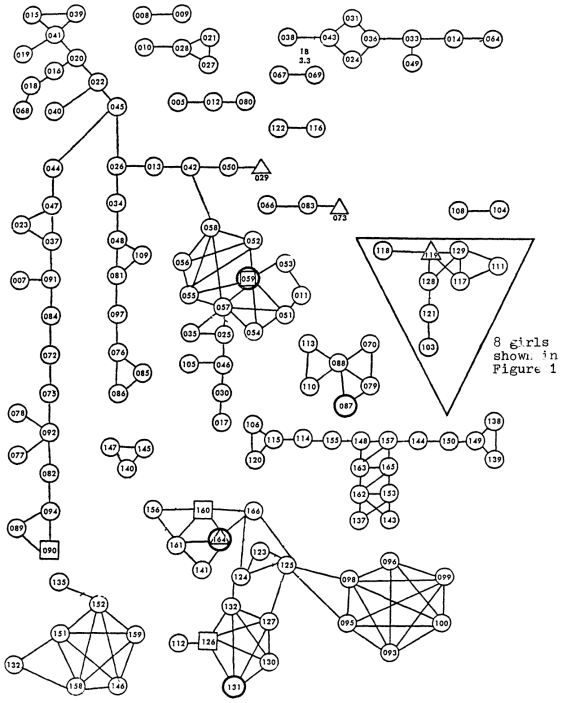
[1]Recall that the variance of a random variable X, is $\mathbb{E}[X^2] - \mathbb{E}[X]^2$. In our case, we can think of degree as a random variable in the following sense. If we draw a node uniformly at random from the graph, we can use X to denote the random variable indicating its degree. If we draw a node i from the graph then its degree is d_i, hence $\mathbb{E}[X] = \sum_i d_i/n$ and $\mathbb{E}[X^2] = \sum_i d_i^2/n$.

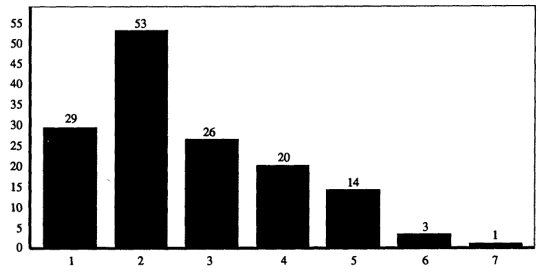
– The friend of T has degree 3.
– The three friends of C have degrees 1, 2 and 2.
– The two friends of L have degrees 2 and 3.
– The two friends of R have degrees 2 and 3.

So indeed, the average number of friends of friends is higher than the average number of friends of individuals. Let's try to understand why this happens. It's because popular nodes contribute disproportionately to the average, since besides having a high score, they're also named as friends more frequently. Watch how this plays out in the example: The score of the node with 1 friend are counted 1 time to compute the average number of friends of friends, whereas the score of nodes with 2 and 3 friends are counted 2 times and 3 times respectively. This is the same kind of weighted average that we have seen in the previous "class size paradox" examples.

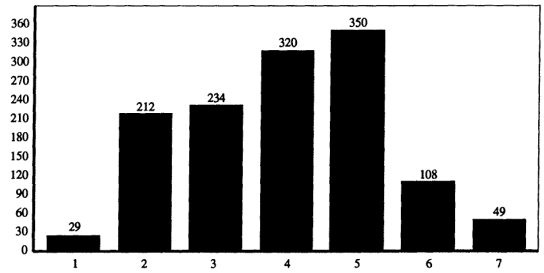
Larger example. In The Adolescent Society, Coleman (1961) collected data on friendships among the students in 12 high schools. Individuals were asked to name their friends, and pairs of individuals who named one another were given particular attention. It is these "friendships" that will be used here.

To illustrate the phenomenon under study here, consider the set of relationships depicted in Figure 2, the complete network of all of the girls in "Marketville", one of the High Schools in the study. Figure 3(a) depicts the distributions of friends of individuals and Figure 3(b) the distribution of friends of friends. Friends tend to have more friends than individuals.





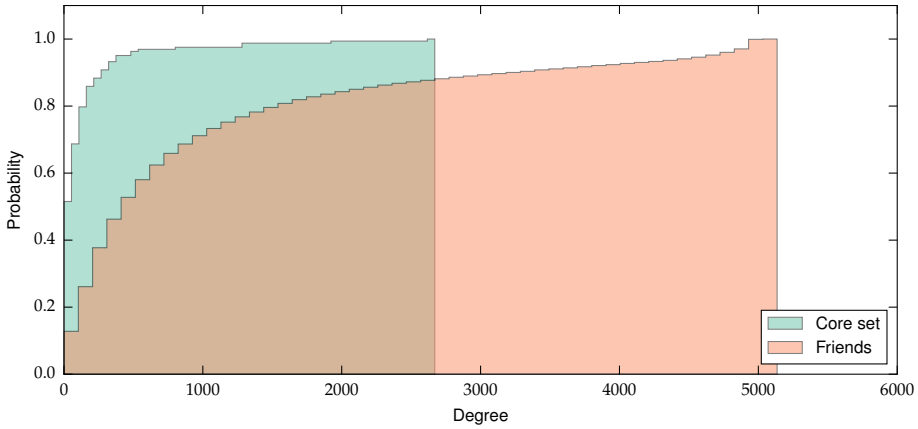
a) The mean is 2.7.

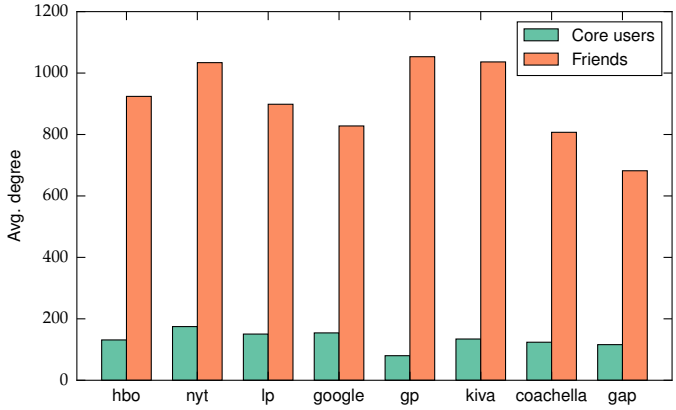


b) The mean is 3.4

(*a*) Distribution of numbers of friends for Marketville girls; (*b*) distribution of number of friends' friends for Marketville girls.

2.1 Friendship paradox in Facebook Pages





3 Survivor Bias: Abraham Wald and the Missing Bullet Holes

The military came to the SRG with some data they thought might be useful. When American planes came back from engagements over Europe, they were covered in bullet holes. Here is the data:

Section of Plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of plane	1.8

The officers saw an opportunity for efficiency; you can get the same protection with less armor if you concentrate the armor on the places with the greatest need, where the planes are getting hit the most. But exactly how much more armor belonged on those parts of the plane? That was the answer they came to Wald for. It wasn't the answer they got.The armor, said Wald, doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines.

Section of Plane Hit	Number of Planes Hit	Number of Planes Back
Engine	20	2
Fuselage	20	9
Fuel system	20	6
Rest of plane	20	10

Wald's insight was simply to ask: where are the missing holes? The ones that would have been all over the engine casing, if the damage had been spread equally all over the plane? Wald was pretty sure he knew. The missing bullet holes were on the missing planes. The reason planes were coming back with fewer hits to the engine is that planes that got hit in the engine weren't coming back. Whereas the large number of planes returning to base with a thoroughly Swiss-cheesed fuselage is pretty strong evidence that hits to the fuselage can (and therefore should) be tolerated. If you go the recovery room at the hospital, you'll see a lot more people with bullet holes in their legs than people with bullet holes in their chests. But that's not because people don't get shot in the chest; it's because the people who get shot in the chest don't recover.

References

[1] Feld, Scott L. *Why your friends have more friends than you do*, American Journal of Sociology, pages 1464–1477, 1991,JSTOR